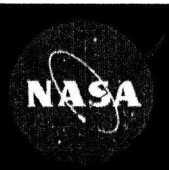


NATIONAL AERONAUTICS
AND SPACE ADMINISTRATION

The Hierarchical Data Format for EOS (HDF-EOS)

Richard Ullman
NASA Goddard Space Flight Center
Greenbelt, MD USA

19th APAN Meetings, Bangkok, Thailand
eScience Workshop, January 27, 2005

What is HDF?

- HDF is a file format and a software library for science data storage, management, exchange, and archiving
  - Highly adaptable, generalized object based data model
  - Completely portable file format - read/write on any computing platform
  - Supports large datasets, as simple or complex in structure as required
  - Designed for high efficiency.
  - Runs on virtually any scientific research computing system
  - Enables a greater degree of collaboration than any other science data format or library available today.
- HDF is written and maintained by the National Center for Supercomputing Applications (NCSA)
  - Rigorous design and testing maintain the library and format for continually evolving scientific computing environments.
  - A highly stable and talented staff are dedicated to assuring that scientific users of HDF receive world-class support.

HDF5 unique and "breakthrough" technology (1)

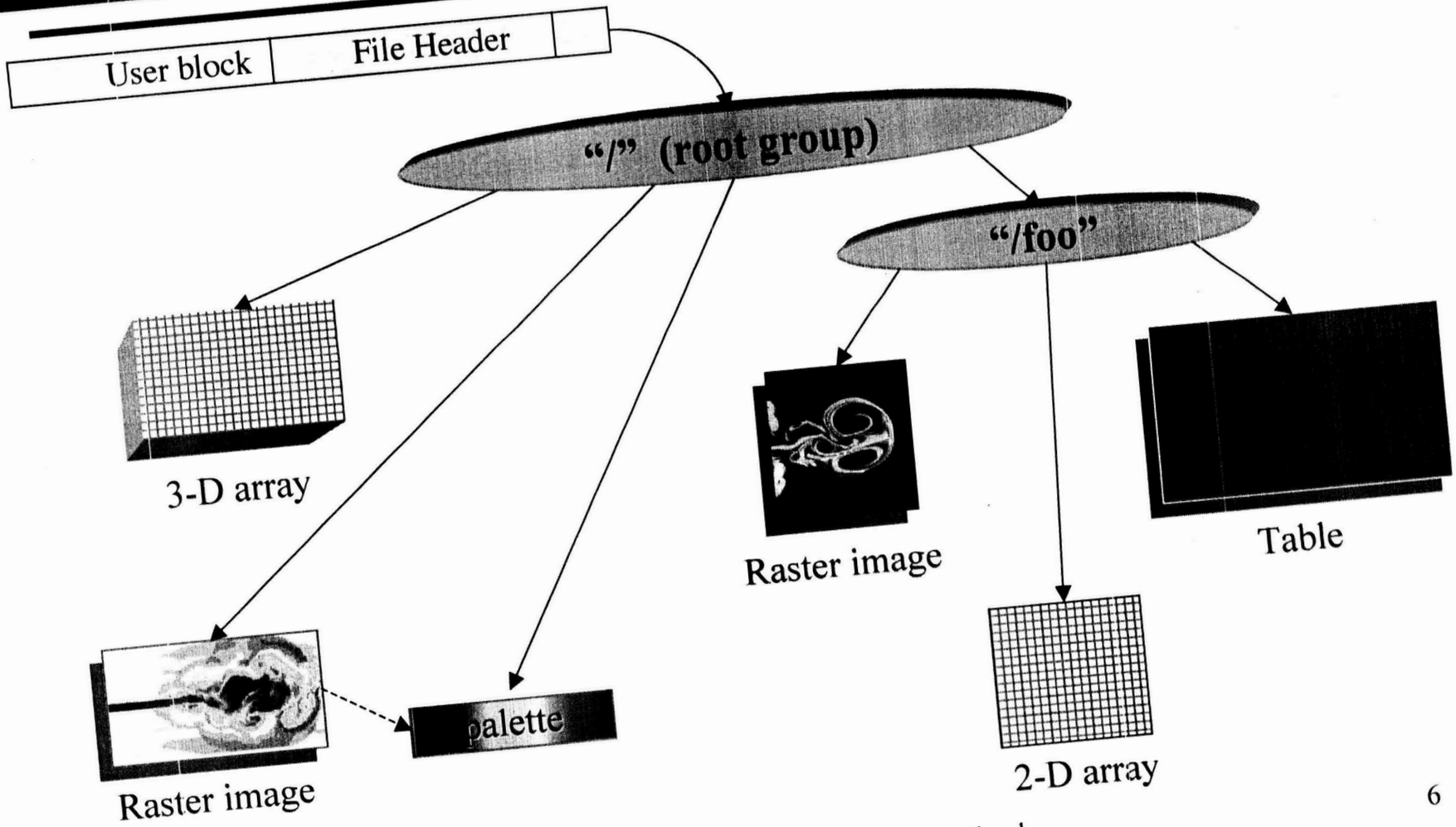
- Unlimited size, extensibility, and portability
  - HDF5 does not limit the size of files or the size or number of objects in a file.
  - The HDF5 format and library are both extensible and designed to evolve gracefully with the articulation of new demands.
  - HDF5 functionality and data is portable across virtually all computing platforms used in scientific research and is distributed with C, C++, Java, and Fortran90 programming interfaces.
- General data model
  - HDF5 has a very simple but versatile data model. The HDF5 data model is compatible most competing formats.
  - Through its grouping and linking mechanisms, the HDF5 data model enables complex data relationships and dependencies.
  - HDF5 accommodates the inclusion of many common types of metadata and arbitrary types and quantities of user-defined metadata.

HDF5 unique and "breakthrough" technology (2)

- Flexible, efficient I/O
  - HDF5, through its virtual file layer (VFL), offers extremely flexible storage and data transfer capabilities by means of special-purpose file configurations and powerful I/O mechanisms, including standard I/O, parallel I/O, and network I/O.
  - An application writer can add additional drivers to implement customized data storage or transport.
  - The parallel I/O driver for HDF5 makes it possible to write data in parallel directly to HDF, resulting in improved access times on parallel systems.
- Flexible data storage
  - HDF5 employs various data compression, data extensibility, and chunking strategies to enhance data access, management, and storage efficiency.
  - HDF5 provides for external storage of raw data, often saving disk space and allowing raw data to be shared among HDF5 files and/or applications.

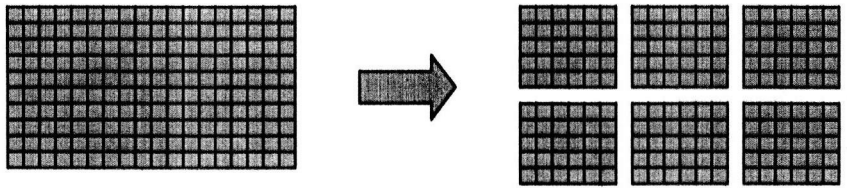
HDF5 unique and "breakthrough" technology (3)

- Unlimited variety of datatypes
  - HDF5 either offers or enables the creation of a virtually unlimited variety of datatypes and imposes no limit on the complexity of a user-defined datatype.
  - Any datatype can be stored in an HDF5 file and shared among other objects in the file, providing a powerful and efficient mechanism for describing data.
  - Datatype storage includes all relevant information, such as endianness, size, and architecture (e.g., IEEE, STD, MIPS).
- Data transformation and complex subsetting
  - HDF5 enables datatype and spatial transformation during I/O operations.
  - HDF5 data I/O functions can operate on selected subsets of the data.


User block
File Header
"/" (root group)
"/foo"
3-D array
Raster image
Table
palette
Raster image
2-D array

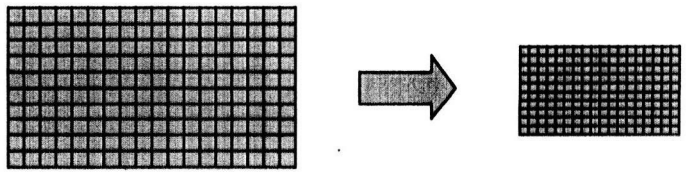
HDF Storage Options

chunked



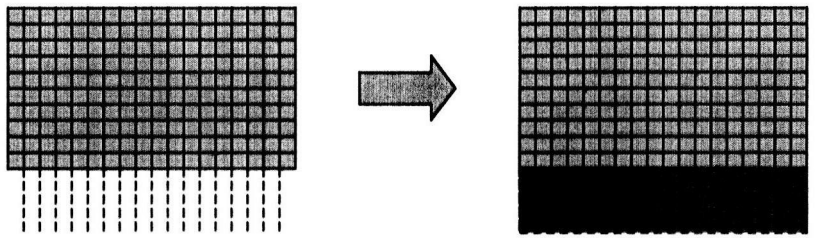
Better subsetting access time; extendable

compressed



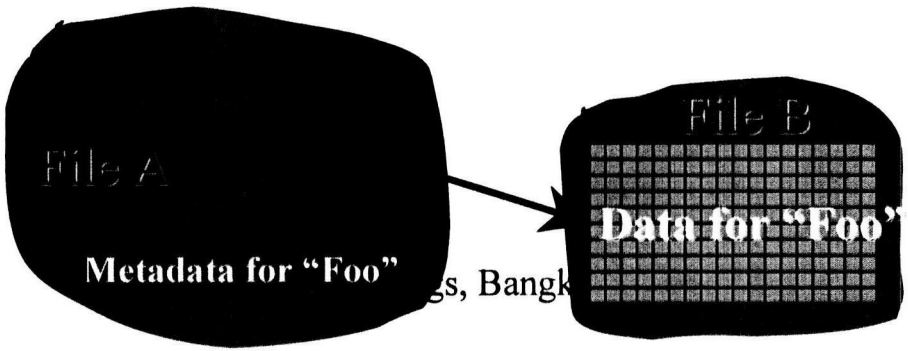
Improves storage efficiency, transmission speed

extendable



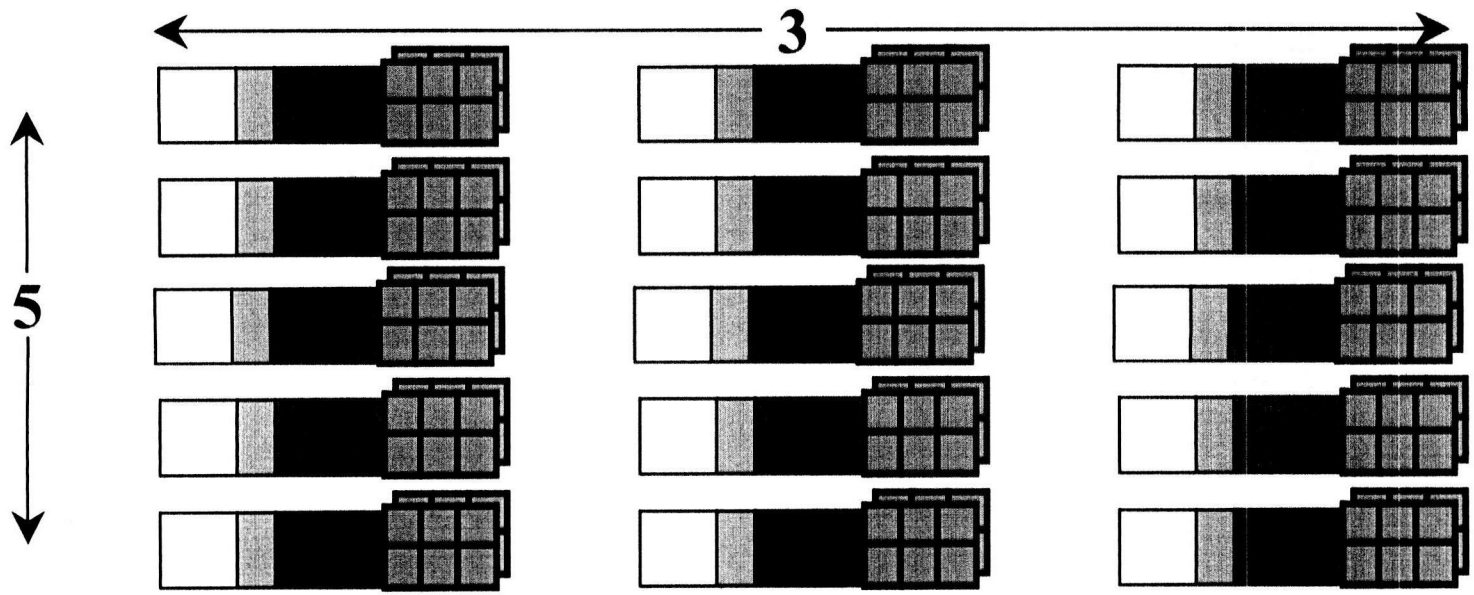
Arrays can be extended in any direction

Split file



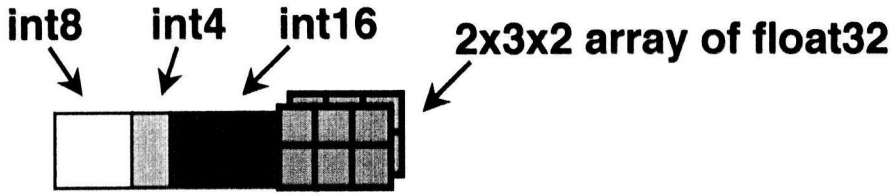
Metadata in one file, raw data in another.

HDF5 dataset: array of records



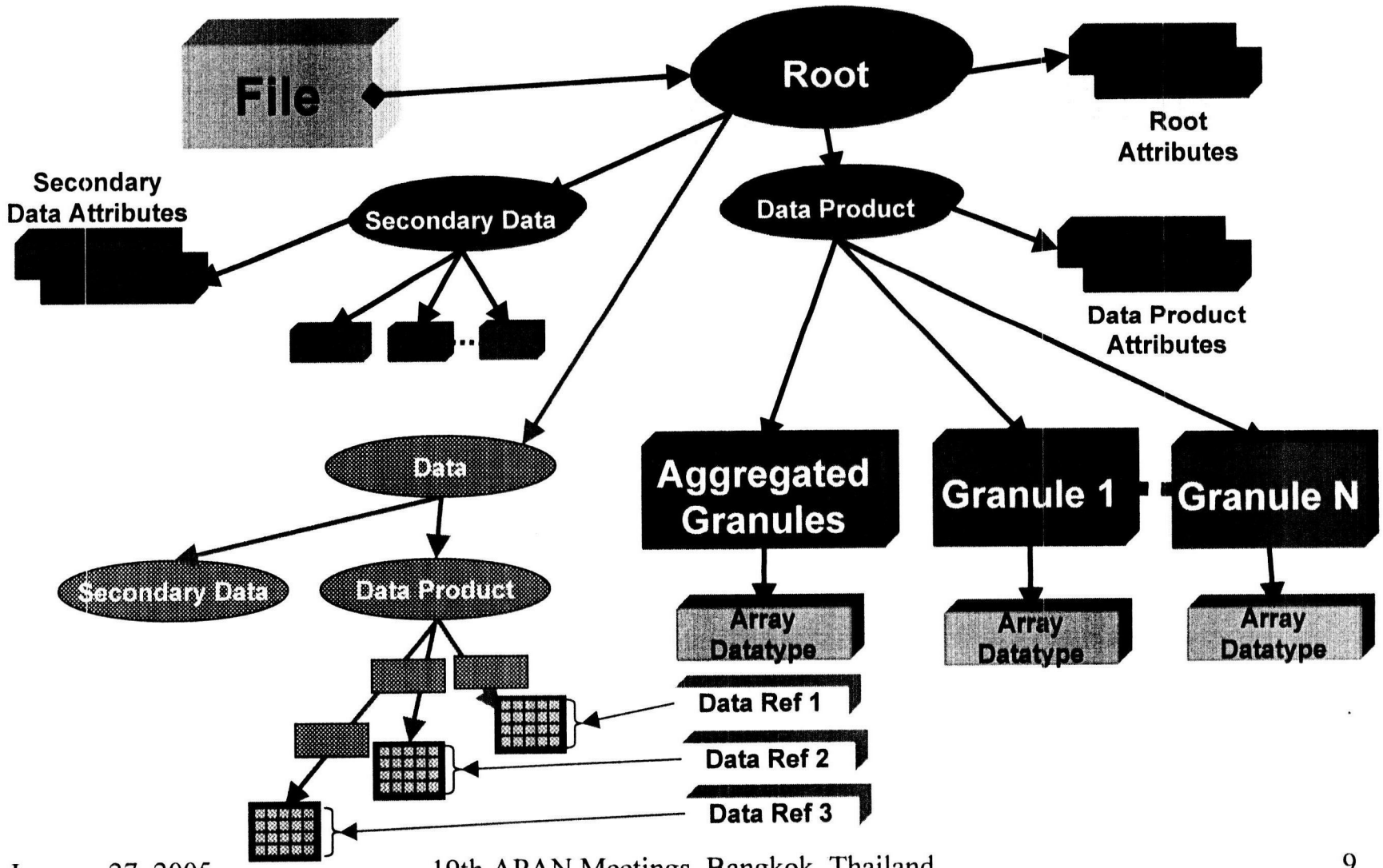
Dimensionality: 5 x 3

Datatype:

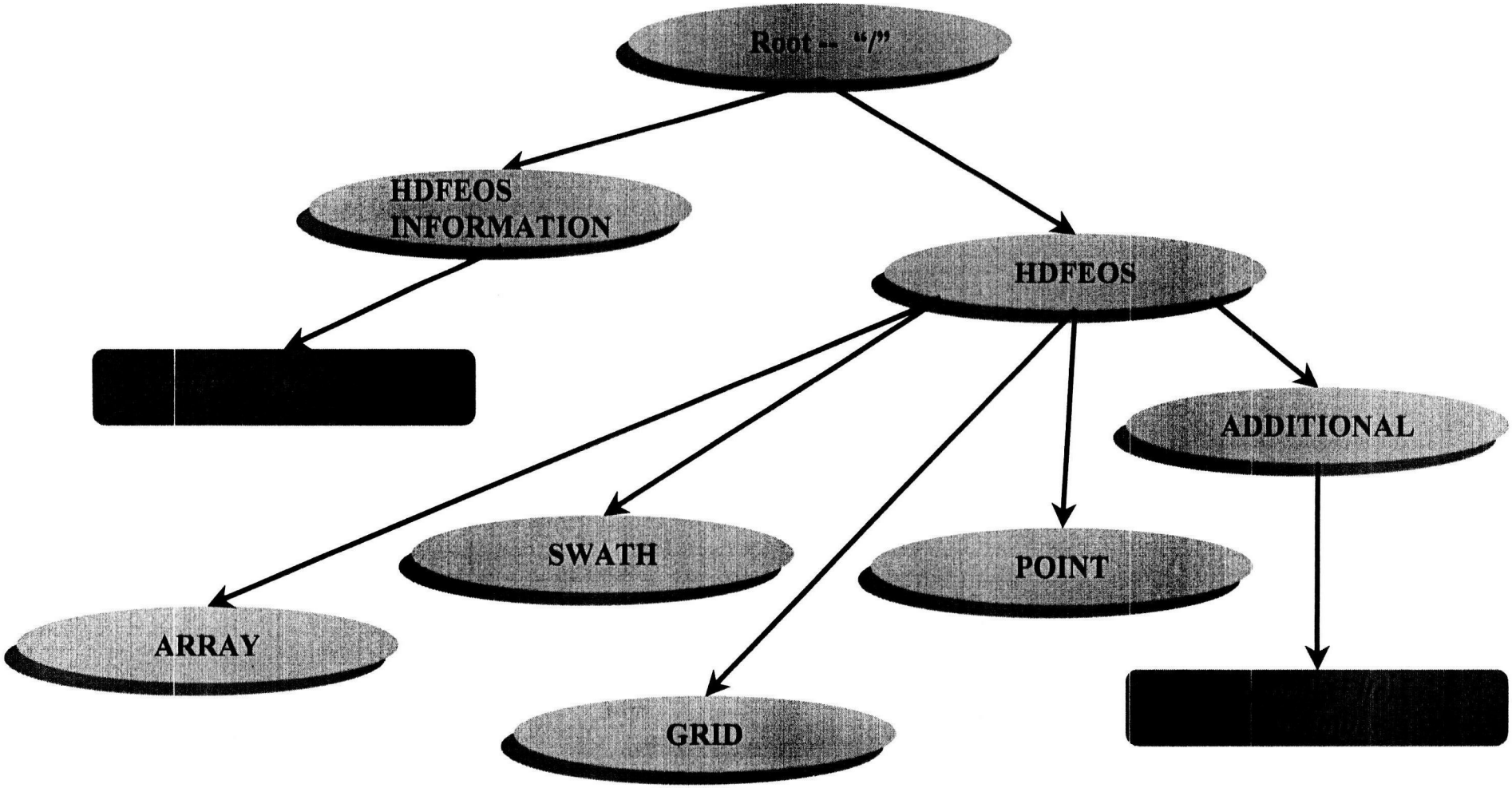


Record

Schematic of an NPOESS HDF5 File

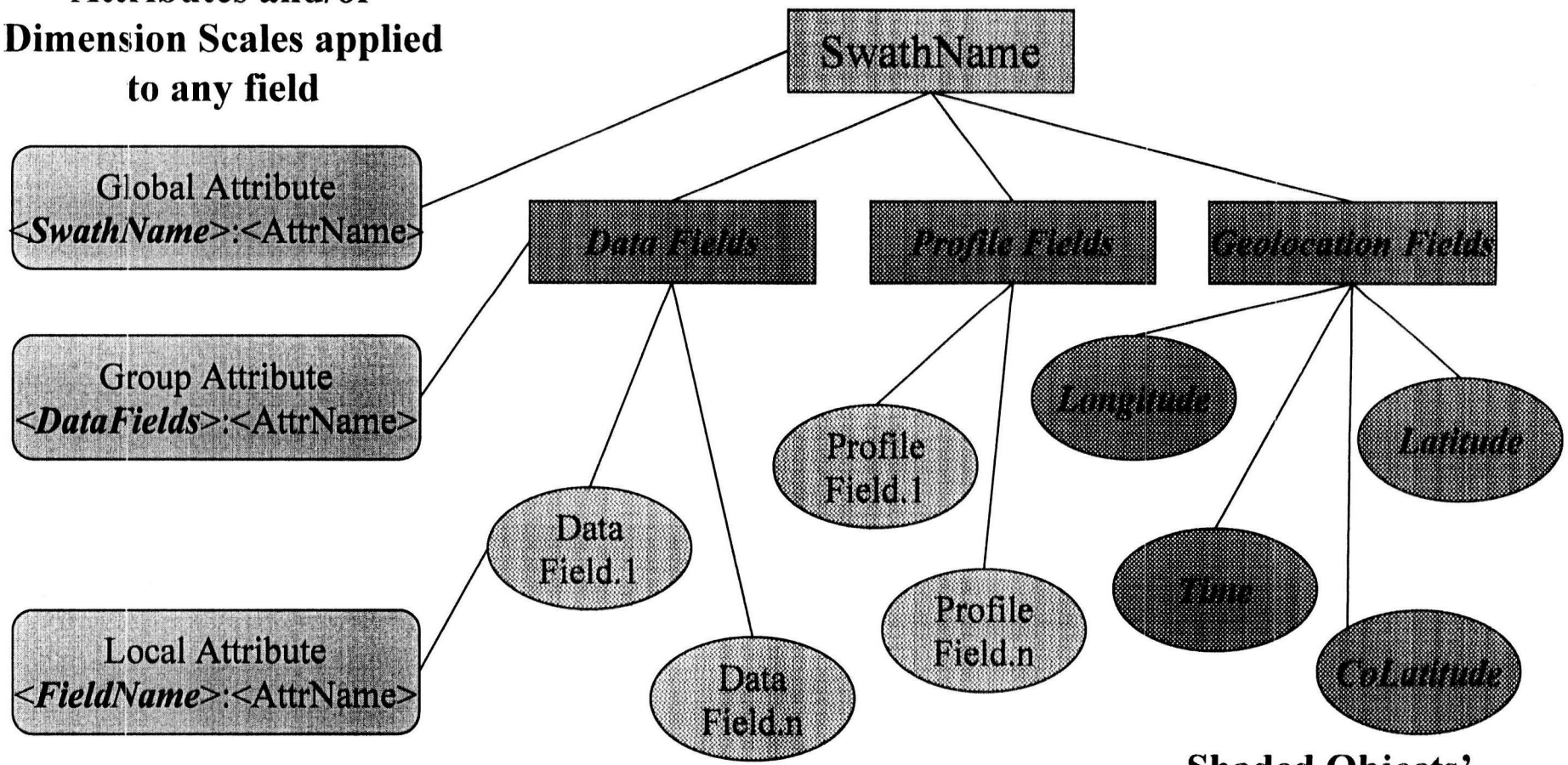


Top Level of an HDF-EOS 5 File



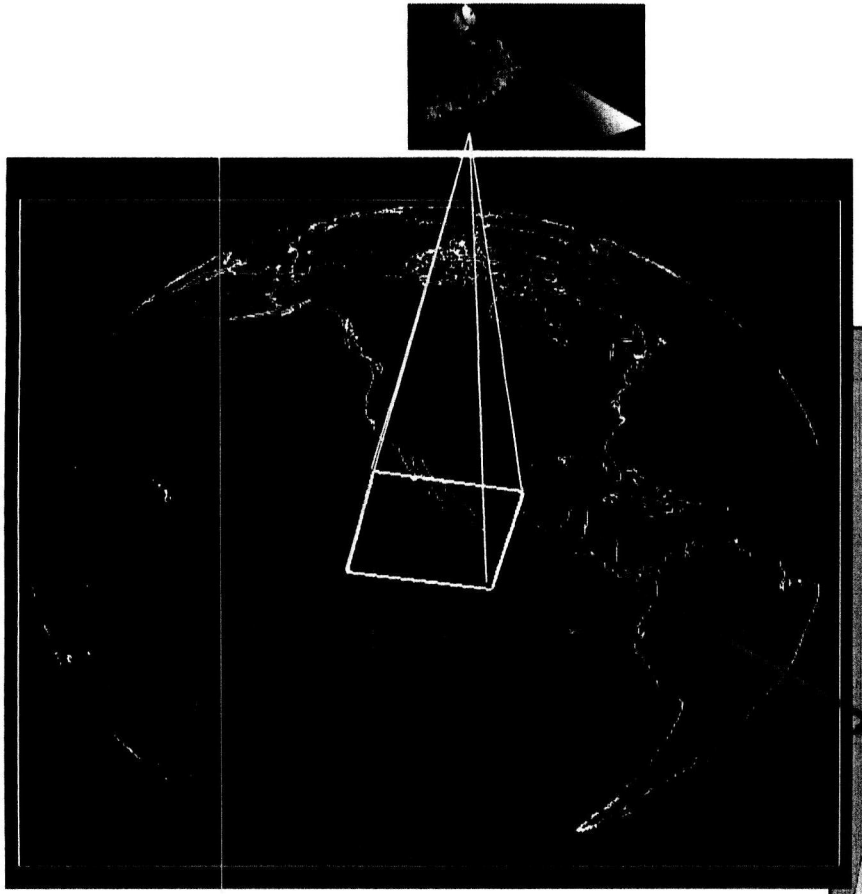
HDF-EOS Swath Structure

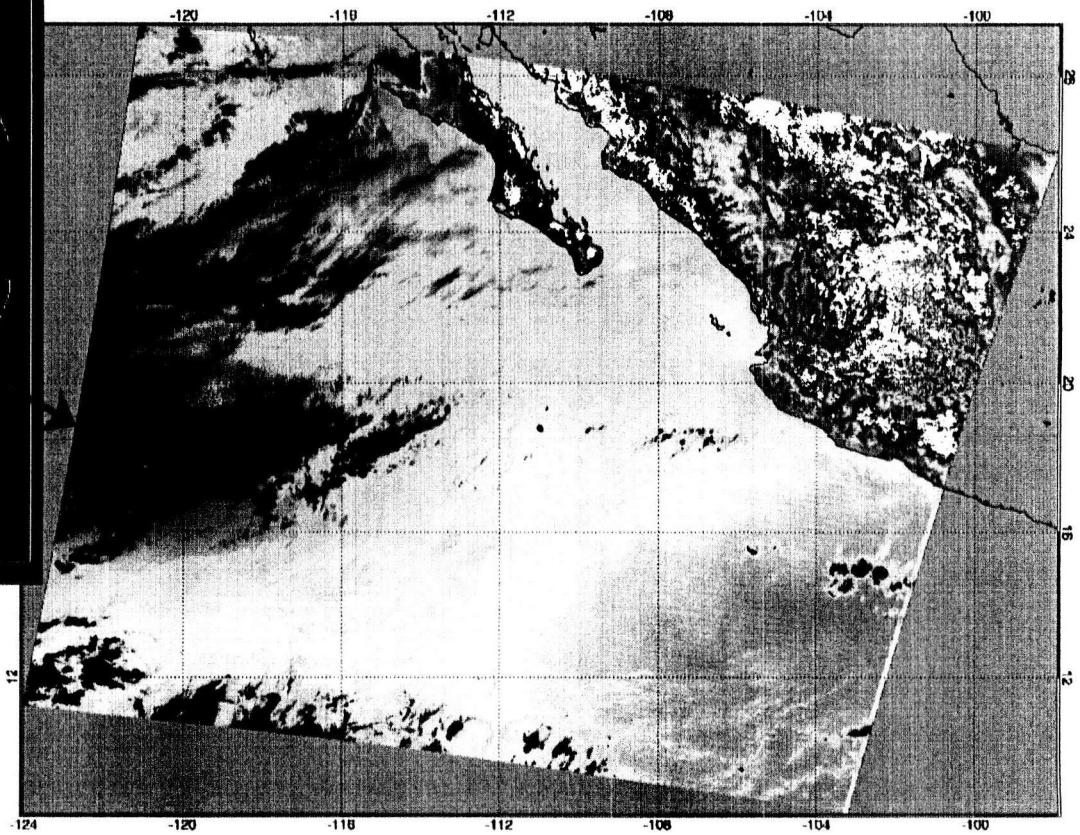
Attributes and/or Dimension Scales applied to any field



Shaded Objects' representation is fixed by library

HDF-EOS Swath application

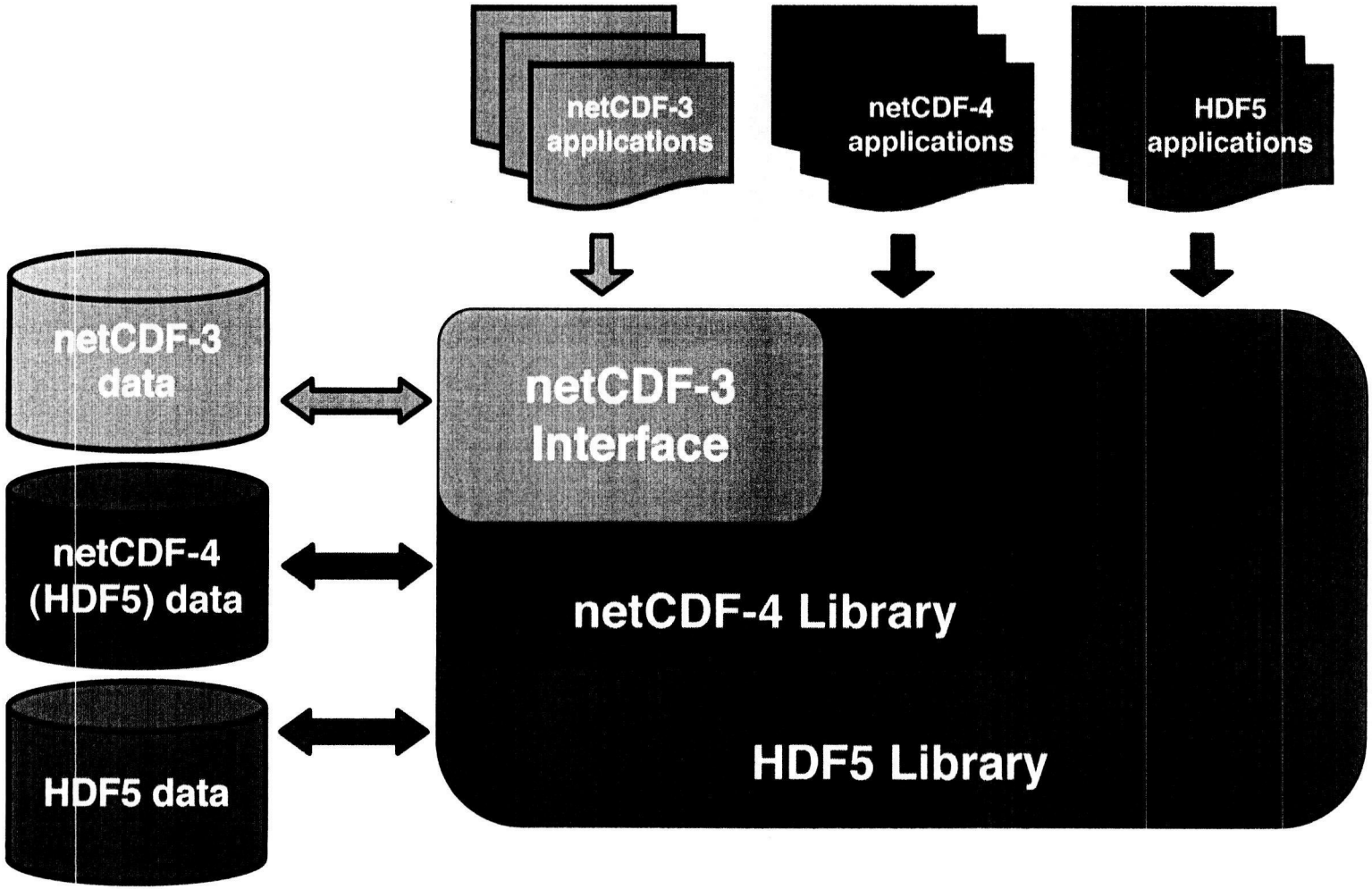




NASA Remote Sensing HDF Data Sources

- Landsat 7 (ETM+)
- Terra (ASTER, CERES, MISR, MODIS, MOPITT)
- Meteor-3M (SAGE III)
- Aqua (AIRS, AMSR-E, AMSU-A, CERES, MODIS)
- Aura (HIRDLS, MLS, OMI, TES)
- OrbView 2 (SeaWIFS)
- TRMM (CERES, PR,TMI, VIRS)
- Quickscat (SeaWinds)
- EO-1 (ALI, Hyperion)
- ICESat (GLAS)
- CALIPSO
- NPP (ATMS, CrIS, OMPS, VIIRS)

netCDF-4 Architecture



Acknowledgements

- Folk, Mike, et al, National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign (UIUC) "HDF5 Nomination for the R&D 100 Award 2002", February, 2002.
- Klein, et al, L-3 Communications Government Services, Inc, "HDF-EOS Development Current Status and Schedule", HDF Workshop, September 2003.
- Rew, Russ, et al, University Corporation for Atmospheric Research Unidata, "Implementing a NetCDF-4 Interface to HDF5 Data", NASA award AIST-02-0071, November 2004.
- Dulaigh, Denise, et al, NPOESS Program, Raytheon Company, Aurora, Colorado"UML Representation of NPOESS Data Products in HDF5", HDF Workshop, October 2004.